HDInsight clusters: Cluster type and configuration

HDInsight clusters, Cluster type and configuration.
This video introduces HDInsight clusters,
the various types of clusters and how to create them.
Then goes over in detail, how to configure them.
The next set of slides provide an overview of how Hadoop
cluster works on HDInsight.
HDInsight provides different types of clusters,
each with a set of components to provide certain functionalities.
The clusters cater to the following uses.
Hadoop for query and analysis using batch jobs.
HBase which provides No-SQL Data storage.
Storm for real-time event processing, and Spark for
in memory processing, interactive queries, and
micro batch stream processing.
Each cluster type has its own terminology for
nodes within the cluster as well as the number of nodes and
the default VM size for each node type.
There are various ways to create HDInsight Linux clusters.
The displayed figure maps the cluster creation options with
the technology that is used for creating the cluster.
This table lists specific information on how to
create a cluster using a method that bests suits your needs.
The Azure Portal is a web based management tool for services and
resources hosted in the Microsoft Azure Cloud.
This section shows how to create Linux based HDInsight clusters
using the portal.
The Core of the current Hadoop version consists primarily
of HDFS in YARN, although MapReduce has historically been
an important service within Hadoop.
As a result, the core cluster configuration files define
the configuration settings for HDFS, YARN, and MapReduce.
These core services are primarily configured using files
installed by Ambari in the ETC Hadoop conf directory.
The core configuration file also define what should be required
to log files, and how to process those log files.
Many Hadoop frameworks, including the core frameworks,
implement Apache Log4j for their logging system.
Hadoop software is installed with a set of default
configuration settings.
These settings are defined in various default.xml files
that are embedded in various JAR files, and are the same for
any installation of the same version of Hadoop.
These default configuration settings are not changed.
Each Hadoop installation includes a set of per
cluster configuration files that commonly include the word site
in their filenames.
For example there is a Core-site.xml,
an Hdfs-site.xml and others.

The per-cluster configuration inherits configuration property settings from the default configuration.
For example, any property that is set in the default configuration is used on the cluster even if that property is not found in the per-cluster configuration.
The per-cluster configuration can also extend the default configuration by adding additional configuration properties.
The per-cluster configuration setting may also override a default configuration setting.
Ambari, either during installation or any time after installation, can be used to modify these files to customize the configuration properties of a specific cluster instance.
Some configuration property settings are based on the hardware configuration of master and worker nodes.
For example, if all of the worker nodes share a common hardware specification, then the same star site.xml files can be used across the entire cluster.
However, if there are two groups of worker nodes with different hardware specifications, then you would need two sets of the star site.xml files, one for each group of worker nodes.
Ambari can manage this for you using a feature named configuration groups which is described later on.
Finally, each job can be submitted with options requesting one or more specific configuration settings.
These requests override the per-cluster and default configuration settings.
For example, the user submitting a job could user the -d option to set a specific configuration property.
This is step 1 of Creating an HDInsight cluster using Azure Portal.
Sign in to the Azure Portal,
click New > Data Analytics > HDInsight.
Type or select the following values.
Cluster name, enter a name for the cluster.
A green check mark will appear next to the cluster name if the name is available.
For cluster types select Hadoop.
Other options include HSpace, Storm and Spark.
For cluster operating system, select Windows.
To create a Linux based cluster, select Linux.
This is step two of creating an HDInsight Cluster using Azure Portal.
Select the appropriate Hadoop version, subscription, select the Azure subscription that will be used for creating this cluster.
Microsoft R Server is the most broadly deployed enterprise class analytics platform for scalable R.
The R language supports a variety of big data statistics, predictive modeling, and machine learning capabilities.

As part of HDInsight Premium,
you can now create an HDInsight cluster with R Server,
ready to be used with massive data sets and models.
This new capability provides data scientists and
statisticians with a familiar R
interface that can scale on demand through HDInsight,
without the overhead of cluster set up and maintenance.
This is step 3, Credentials.
Configure the username and password for
the Hadoop cluster HTTP user.
If you enable remote desktop for the cluster,
you will need to configure the remote desktop user username and
password and an account expiration date.
Click Select at the bottom to save the changes.
WASB or ADLS storage options could be used depending on
the type of cluster chosen.
Step 4, selecting WASB as the data source.
Create a new or select an existing Azure Storage account
to be used as the default file system for the cluster.
Selecting ADLS as the data source.
In the data source blade, select the details for
the storage account and storage container.
On the cluster AAD identity blade,
you can choose to select an existing service principle or
create a new one, then click Create.
Data source.
On the cluster AAD identity blade, click Manage ADLS access.
The pane shows the Data Lakes Store accounts associated with
the subscription.
However, you can set the permissions only for
the account that you created.
Select Read/write execute permission for
the accounts you want to associate with the HDInsight
cluster, and then click Save Permissions.
Now, specify your cluster configuration.
In the Node Pricing Tiers,
set the number of worker nodes that you need for the cluster.
The estimated cost of the cluster will be shown within
the blade.
To select the cluster version as well as configure other
optional settings, such as joining a virtual network or
setting up an external metastore to hold data for
Hive and UZI, use script actions to customize a cluster,
to install custom components or
use additional storage accounts with the cluster, click Create.
Selecting Pin to Start Board will add a tile for
the cluster to the start board of your portal.
The icon will indicate that the cluster is being created and
will change to display the HDInsight icon
once creation has completed.

It will take some time for
the cluster to be created, usually around 15 minutes.
Use the tile on the Start Board, or
the Notifications entry on the left side of the page,
to check on the provisioning process.